基于动态层次 K－Modes 的电网数据聚类分析

林红阳[1]，杜　翼[1]，刘　林[1]，易　杨[1]，蔡　菁[2]，马汉斌[2]
（1. 国网福建省电力有限公司经济技术研究院，福建 福州　350012；
2. 国网信通亿力科技有限责任公司，福建 福州　350000）

摘　要：随着电力工业的发展与电力计量体系的不断完善，电网需求侧用电特性呈多样化发展态势。智能电表走进人们的生活，带来海量电力数据。如何挖掘用户的用电行为特性，从而促进电价市场化成为人们所关心的问题。首先介绍了聚类分析中的 K－modes 算法以及层次 K－means 算法，并结合考虑其优缺点提出动态层次 K－modes 算法来处理类属型数据并给出合理的 k 值；其次提出了将曲线数据进行差分及类属型转化的数据处理方法，使之能更好地反应用户曲线形态；最后利用动态层次 K－modes 算法在模拟数据以及厦门岛内地区电力用户的真实数据上进行聚类试验，得到优良的分类结果。

关键词：电网系统；聚类分析；动态层次 K－modes 算法；曲线数据；类属型转化
中图分类号：TP181　**文献标志码**：A　**文章编号**：1003－6954（2019）06－0030－06
DOI:10.16527/j.cnki.cn51-1315/tm.2019.06.007

Clustering Analysis of Power Grid Data Based on Moving Hierarchical K－modes Method

Lin Hongyang[1], Du Yi[1], Liu Lin[1], Yi Yang[1], Cai Jing[2], Ma Hanbin[2]
(1. Institute of Economics and Technology, State Grid Fujian Electric Power Co., Ltd., Fuzhou 350012, Fujian, China;
2. State Grid Telecommunication Yili Technology Co., Ltd., Fuzhou 350000, Fujian, China)

Abstract: Withing the development of power industry and the continuous improvement of electric power metering system, the features of power grid demand side present a diversified development. Smart meters come into people's life, which is bringing huge amounts of electricity data. A meaningful problem is how to explore the behavior of users to promote the electricity market. Combining the considerations of traditional K－modes method and hierarchical K－means method, moving hierarchical K－modes method is proposed, which is able to deal with categorical data and gives a reasonable k value of the number of clusters. Moreover, a method to transform curve data into categorical data by differencing and categorizing is proposed, which can preferably reflect the shape of curves. At last, clusetering experiment is carried out by using moving hierarchical K－modes method based on simulated data and real data of power users in Xiamen island region, which gets excellent clustering results.
Key words: power system; clustering analysis; moving hierarchical K－modes method; curve data; categorizing

随着电力工业的发展与电力计量体系的不断完善，在电力用户侧积累了海量的历史数据。针对既有数据，展开用户用电行为特性的挖掘，可以为电力公司开展相应电价制定与需求侧管理工作提供有益指导[1]。聚类分析是数据挖掘（data mining）的重要组成部分，作为一种无监督学习方法，根据其思想的不同，聚类算法主要可分为以下5种方法：划分法（partitioning method）、层次法（hierarchical method）、基于密度法（density－based method）、基于网格的方法（grid－based method）、基于模型法（model－based method），并且在此基础上发展出更为灵活的组合及变形。而 K－means 算法作为一种基础的划分法，从提出以来就受到人们的普遍关注，至今仍有不少学者对 K－means 及其同类型方法在应用及算法上做出研究贡献[2－11]。K－modes 作为 K－means 的一种变形，能够对类属型数据进行分类，其主要区别在于中心和距离的定义[12]。而层次聚类法作为另一种广泛被人们使用的算法，其优势在于分类准确

且可以生成直观的分类树，便于判断类的数目；但其计算量过大，算法复杂度至少为 $O(n^2)$，仅适用于小规模数据聚类。也有将两种或多种传统聚类算法结合而成的新算法，例如层次 K－means 算法[3,11]。此类算法能够保留每种算法的优良性并弥补算法缺陷，使得算法的适应性和准确性都得到提升。下面通过引入合适的聚类算法对厦门岛内地区电力用户数据进行试验，以验证算法的可行性，为进一步推广至其他数据集筹备基础。

1 算法介绍

1.1 K－modes 算法

作为一种被普遍运用于各类问题的聚类算法，K－means 常用于数值型数据的分类，距离一般采用欧氏距离，所以对其他类型数据，例如类属型（0－1 型）数据并不适用。作为改进，由 Huang 在 1998 年提出的 K－modes 算法在 K－means 的算法基础上引入 modes 取代 means（中心），并提出差异匹配（matching dissimilarity）替代传统的欧氏度量（距离）。正是引入差异度量（dissimilarity measure）这种距离计算方式，使得其能对类属型数据进行有效的划分。

K－modes 算法从构造思想上与 K－means 基本一致[6,9]。定义信息集合 (X,A,Q)，其中 $X=\{x_i \mid i=1,2,\cdots,n\}$ 表示数据集合；$A=\{a_i \mid i=1,2,\cdots,m\}$ 表示数据每一维度的属性；$Q=\{q_i \mid i=1,2,\cdots,k\}$ 表示 k 个分类，而 $\{q_i \mid i=1,2,\cdots,k\}$ 表示每个类的中心（mode）。其中 $x_{ij} \in z$，$(i=1,2,\cdots,n; j=1,2,\cdots,m)$ 表示样本点 x_i 在 a_j 属性上的取值。

定义任意两点 x，y 的差异度量为 d：

$$d(x,y)=\sum_{z=1}^{m}\delta(x_z,y_z)$$

其中，

$$\delta(x_z,y_z)=\begin{cases}1(x_z=y_z)\\0(x_z\neq y_z)\end{cases}$$

定义目标函数为

$$F(X,Q)=\sum_{j=1}^{k}\sum_{i=1}^{n}w_{ij}d(x_i,q_j)$$

其中，

$$w_{ij}\in\{0,1\}\quad i=1,2,\cdots,n; j=1,2,\cdots,k$$

$$\sum_{j=1}^{k}w_{ij}=1\quad i=1,2,\cdots,n$$

$$0<\sum_{i=1}^{n}w_{ij}<n\quad j=1,2,\cdots,k$$

在上述 3 个条件下使目标函数 F 达到最小值，K－modes 基本实现步骤如下：

1）随机选取 k 个 modes。

2）计算每个数据点与 k 个 modes 的差异度量 d（距离），将每个数据点分入度量值最小的类。

3）判断是否达到迭代终止条件（一般设置终止条件为分类结果不再改变），若终止迭代则返回结果，否则进入步骤 4）。

4）对每一类中所有数据点在每一维度上取众数，更新每个类的 mode：$q_{zj}=\mathrm{mode}(x_{ij} \mid x_i \in Q_z)$，$j=1,2,\cdots,m$，返回步骤 2）。

1.2 层次 K－means 算法

K－means 聚类算法复杂度为 $O(n)$ 阶，即不需要进行样本点的两两距离计算，在处理一般聚类问题时速度较快。但 K－means 算法也存在 2 个公开问题：1）类数目无法自适应得到；2）算法仅能保证收敛到局部最优解。而层次聚类算法能够得到较好的聚类结果，同时可以通过聚类系谱图来指导确定类的数目，但由于其算法复杂度过大，当处理大规模数据时就会遇到困难。

由 Arai 于 2007 年提出的层次 K－means[3] 从一定程度上解决了这些问题。其利用所有在特定 k 值下运行 K－means 所得到的结果，这些结果可能均收敛到局部最优解，但通过大量结果所带来的信息，结合层次聚类算法对其进行转换便能确定出更为准确的 K－means 初始值中心点。

层次 K－means 算法主要步骤如下：

1）取定 k 值，并进行 p 次重复的 K－means 计算，得到聚类中心点集合；

2）结合层次聚类法，对步骤 1）所得的聚类中心点集合进行再聚类；

3）将步骤 2）所得的聚类中心点作为 K－means 的初始聚类中心点，并运行 K－means 算法得到最终聚类结果。

1.3 层次 K－modes 算法

考虑到 K－modes 算法对类属型数据的适应性，以及其在迭代过程中与 K－means 算法的相似性，因此提出层次 K－modes 算法。将 K－modes 算法与层次聚类算法进行结合，继承两种算法的优点

而直接作用于类属型数据。值得注意的一点是差异度量(dissimilarity measure) 的计算快于同等维度的二范数计算,因此从速度上层次 K – modes 也比层次 K – means 更具优势。其算法实步骤与层次 K – means 类似,不过类中心必须使用 mode,并且在层次聚类时也必须选取适合类属型数据的距离公式进行计算。

1.4 动态层次 K – modes 算法

进一步,还希望借助聚类系谱图来选取适合的 k 值。层次 K – modes 中,由于固定了 k 值,因此进行 p 次相同 k 值的 K – modes 并完成层次聚类后所生成的系谱图所呈现的划分很大程度上受到所选取 k 的影响,通常与初始选取的 k 值相同。为了克服这个弱点,提出让每次 K – modes 的 k 值变动起来,从而弱化人为选取 k 值的影响[14]。在这种改动下,p 次 K – modes 带来更多的分类信息,从而能从更广范围内寻找更优的 k 值。

动态层次 K – modes 主要算法步骤如下:

1) 将 k 值分别设置为从 2 至 $p+1$ 的整数,进行 k 次 K – modes 聚类,得到每一次的类 mode,将 p 次 K – modes 共有 $\sum_{i=1}^{p} k_i$ 个 modes 保存为集合 M;

2) 利用层次聚类法对集合 M 中的 modes 进行分类;

3) 通过系谱图选定适当的 k 值以及对应的 modes;

4) 利用步骤 3) 中结果作为初始条件进行一次 K – modes,并返回结果。

2 数据特征提取

将曲线数据按照曲线形态进行聚类,考虑到直接利用原始数据进行聚类会忽略掉时间序列上段与段间的形态差异,造成聚类结果不合理。

如文献[13]将原始数据进行平滑后取一阶差分值,将原始矩阵 X 转化为差分矩阵 D,其每一行为 d_i,由一个 m 维行向量构成,表示第 i 个差分后样本;更进一步,分别将“$\pm 0.1 \times$差分样本极差”作为阈值 $t(d_i)$。

$$ca_{ij}=\begin{cases}1, d_{ij}>t(d_i)\\ 0, |d_{ij}|<t(d_i)\\ -1, d_{ij}<-t(d_i)\end{cases} \quad (1)$$

式中,$j=1,2,\cdots,m$。

对所有差分样本进行式(1) 处理后得到类属型矩阵 $\boldsymbol{C}$,为 $n \times m$ 维的矩阵。其每一行为 c_i,由一个 m 维行向量构成,以此来反映曲线形态。

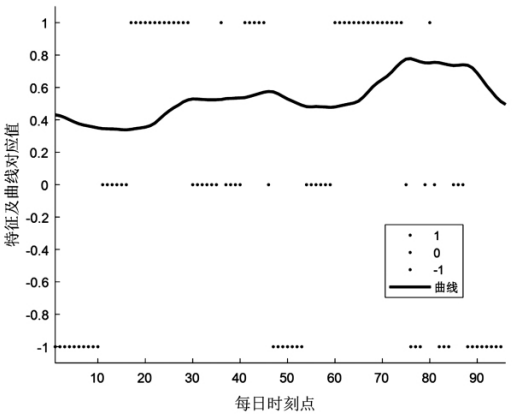


图 1 特征提取

图 1 为数据特征提取图,图中的曲线为某样本 x_i 在经过平滑后得到,经过差分及式(1) 处理后得到的圆点,表示在每一时刻该样本曲线的升、平、降 3 种状态,将连续型的时间序列数据转化成类属型的离散状态值,从而提取出样本的形态特征,可利用动态层次 K – modes 算法进行聚类。

3 算法实验

3.1 模拟数据试验

为了检验动态层次 K – modes 相较于传统 K – modes 的优良性,首先使用计算机模拟出一个维度为 10,类数目为 3,类内样本量分别为 500、200、50 的曲线数据集,如图 2 至图 4 所示,其中双峰形类样本数为 500,单峰型类样本数为 200,单谷型类样本数为 50。

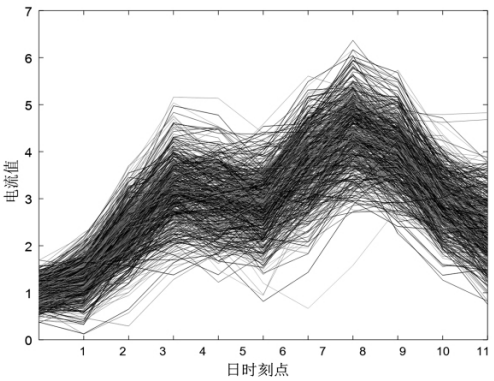


图 2 双峰型类

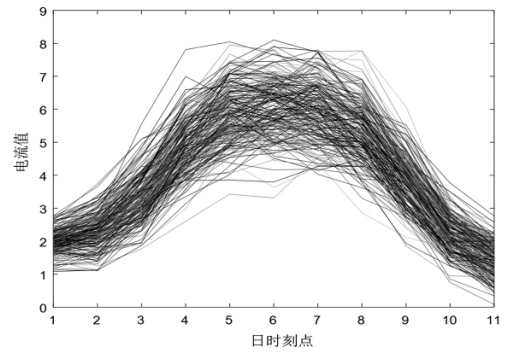


图 3　单峰型类

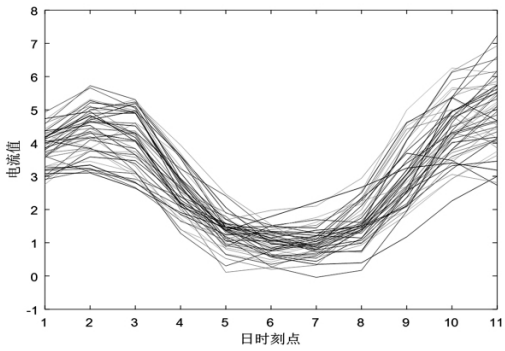


图 4　单谷型类

进行算法比较前，首先给出分类准确率的计算规则为

$$\mathrm{CE}=\frac{\sum_{i=1}^{k} n_i}{N}\times 100\%$$

式中，n_i为当前聚类结果中第 i 类包含正确分类中最多一类的数目。

对于传统的 K－modes 算法而言，k 必须经由人工进行初始化。为了进行更客观的比较，假定已知类数目为 $k=3$，并利用 K－modes 算法对该模拟数据进行聚类，聚类结果如图 5 所示。

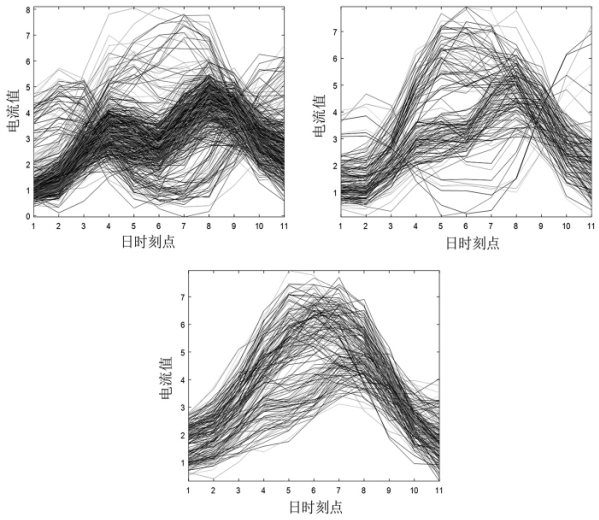


图 5　普通 K－modes 聚类(k=3)

如图 5 中各子图以及图 2 至图 4，分类准确度仅达到 CE=47.87%，即该聚类结果大部分类都未能准确对应正确分类情况。

而使用动态层次 K－modes 算法对此模拟数据进行聚类，并不需要人为取定 k 值，而仅需通过返回的系谱图来决定所希望的 k 值以及初始 modes。取 K－modes 运行次数 $p=8$，因此 8 次计算的 k 值分别为 2～9。在层次聚类时选择“hamming”距离，应用“average”的类间距计算方式，得到如图 6 所示系谱图。

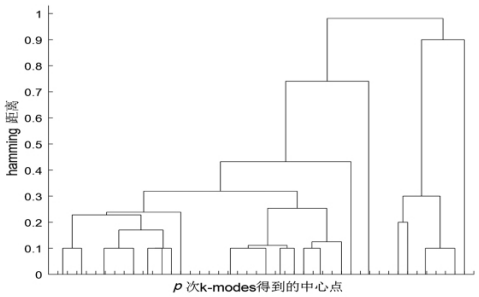


图 6　聚类系谱

从图 6 中可见，聚类树在 $k=3$ 时的高度差最为明显，可从 $k=3$ 处进行截取，同时返回对应的中心(modes)。在此基础上进行一次 K－modes 可得到如图 7 所示层次 K－Modes 聚类图。

如图 7 所示，在同样 $k=3$ 的情形下，动态层次 K－modes 相较于传统 K－modes，分类准确度达到 CE=76.93%，明显高于传统 K－modes；并且类数目 k 也能被正确确定，因此其分类效果明显优于传统 K－modes。

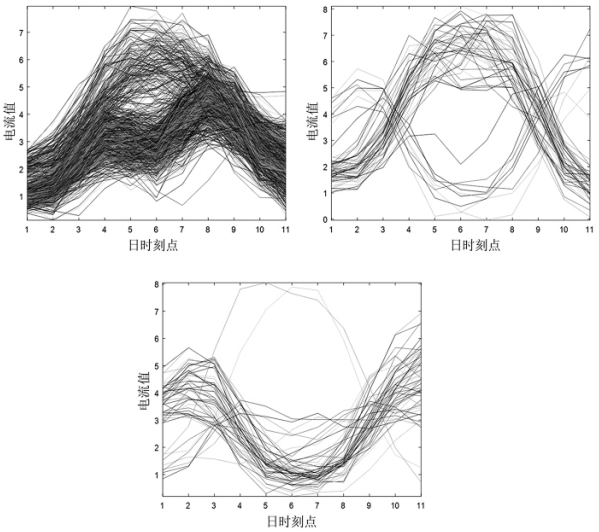


图 7　层次 K－modes 聚类

3.2　真实数据试验

所采用的真实数据为来自厦门某地区 2016 年 3 月 1 日至 4 月 13 日的用户电流数据，为曲线数

据。数据规模为 44 × 9026 × 96，每一维度含义分别为：采集天数为 44，用户变压器数目为 9026，每日共 96 个观测值（每日隔 15 min 一次观测）。

考虑到用户用电曲线基本以日为周期，因此对 44 天内所有用户的每日 96 次观测分别取平均值，将数据降维，得到矩阵 $\boldsymbol{X}$（9026 × 96），其中矩阵每行 x_i 为一个样本，由一个 96 维行向量构成。

对 X 运用式（1）的数据处理方法，将其转化成类属型数据矩阵，并利用动态层次 K - modes 对数据进行分类。取 K - modes 的运行次数 $p = 8$，在层次聚类时同样选择“hamming”距离，类间距计算方式选择“average”得到聚类系谱图，如图 8 所示。

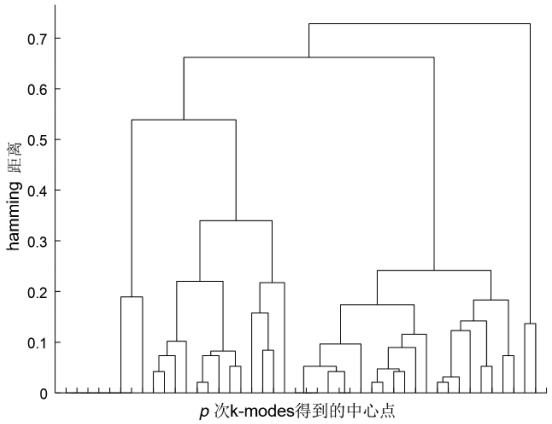
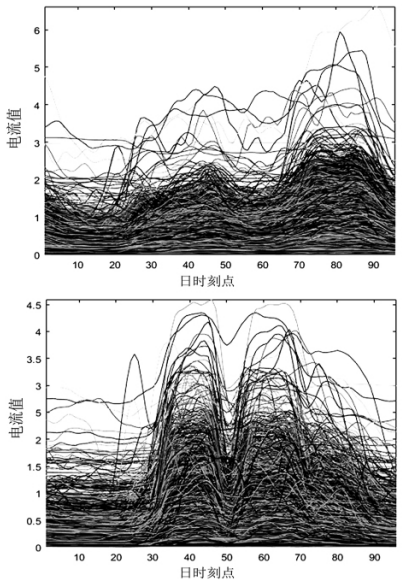


图 8　中心层次聚类系谱

对图 8 聚类树进行水平截断，具有明显高度差的分类情形为 $k = 3$、$k = 6$。因此分别选取 $k = 3$ 和 k = 6 的情形，给出中心（modes）作为初始值，进行 K - modes 聚类。

在 k 值为 3 的情形中，得到图 9 所示分类结果。



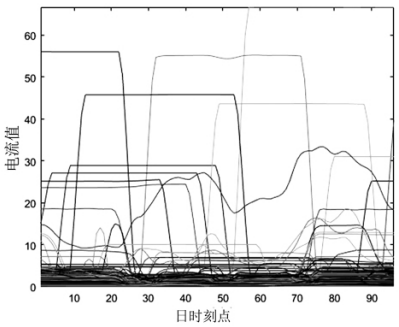


图 9　层次 K - modes 聚类（$k = 3$）

如图 9 中 3 个子图所示，前两个类特征明显，属于双峰型类，但峰型却有差异；第一个类双峰集中在 30 ~ 70 时段内，第二类中两个峰在分布在 24 ~ 50 以及 65 至 85 时段内。在 $k = 3$ 的情形中，此算法抓住了曲线数据的主要形态特征给出合理的分类结果。

在 k 值为 6 的情形中，得到如图 10 所示的分类结果。

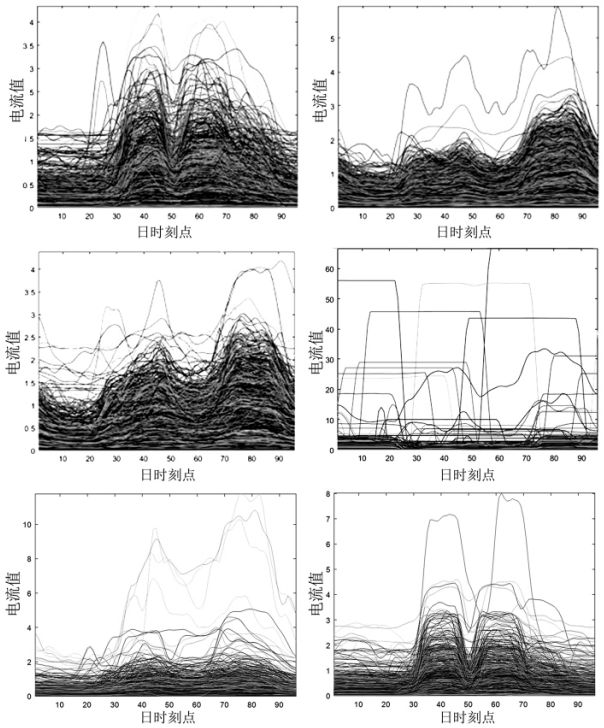


图 10　层次 K - modes 聚类（$k = 6$）

相较下，图 10 所示的各类峰型均表现出较明显的差异。相较于模拟数据，真实数据表现出了更复杂的分布特性。正是由于这种复杂性，很难简单给出唯一的分类结果；正如图 8 中聚类树的枝干在 $k = 3$、$k = 6$ 时都有明显分化。

4　结　语

传统 K - modes 算法与层次聚类算法结合，成

功地将层次 K – means 的算法优点移植到类属型数据中。并且所提出的动态层次 K – modes 算法还可以通过聚类系谱图确定 k 值，一定程度上解决了 k 值须初始给定的公开问题。算法也具有更强的适应性和异常点识别等优良性质，可以对真实数据给出有效聚类结果。

考虑到动态层次 K – modes 算法从结果上虽然明显优于传统 K – modes 算法，但在进行聚类时仍会出现一定的错分现象，在后续工作中考虑引入 robust 思想，无须将所有样本带入聚类迭代过程，而是仅选取一部分重要样本进行迭代，以此提高算法准确度。同时在特征提取方面，也希望引入更精细化的自适应阈值给定，从数据角度提高聚类质量。

参考文献

[1] 赵岩，李磊，刘俊勇，等. 上海电网需求侧负荷模式的组合识别模型[J]. 电网技术，2010，34(1)：145 – 151.

[2] ZheXue Huang. Extensions to the k – Means Algorithm for Clustering Large Data Sets with Categorical Values [J]. Data Mining and Knowledge Discovery，1998(2)：283 – 304.

[3] Kohei Arai，Ali Ridho Barakbah. Hierarchical K – means: An Algorithm for Centroids Initialization for K – means [N]. Reports of the Faculty of Science and Engineering，Saga University，2007，36(1)：25 – 31.

[4] 金建国. 聚类方法综述[J]. 计算机科学，2014，41(11A)：288 – 293.

[5] Deng Hai，Tan Hua，Sun Xin. A K – Means Clustering Algorithm of Meliorated Initial Center [J]. Computer Technology and Development，2013，23(11)：42 – 45.

[6] 苏锦旗，薛惠锋，詹海亮. 基于划分的 K – 均值初始聚类中心优化算法[J]. 微电子学与计算机，2009，26(1)：8 – 11.

[7] Jinhua Li，Shiji Song，Yuli Zhang，et al. Robust K – Median and K – Means Clustering Algorithms for Incomplete Data [J]. Mathematical Problems in Engineering，2016：1 – 8.

[8] Muhammad Umer Munir，Muhammad Younus Javed，Shoab Ahmad Khan. A Hierarchical k – Means Clustering Based Fingerprint Quality Classification [J]. Neurocomputing，2012，(85)：62 – 67.

[9] Caiming Zhong，Duoqian Miao，Pasi Frnti. Minimum Spanning Tree Based Split – and – merge：A Hierarchical Clustering Method [J]. Information Sciences，2011，16(181)：3397 – 3410.

[10] Preeti Aroraa，Deepali Dr. b，Shipra Varshneyc. Analysis of K – Means and K – Medoids Algorithm for Big Data [J]. Procedia Computer Science，2016(78)：507 – 512.

[11] Kaiyang Liao，Guizhong Liu，Li Xiao，et al. A Sample – based Hierarchical Adaptive K – means Clustering Method for Large – scale Video Retrieval [J]. Knowledge – Based Systems，2013(49)：123 – 133.

[12] Liang Jiye，Bai Liang，Cao Fuyuan. K – Modes Clustering Algorithm Based on A New Distance Measure [J]. Journal of Computer Research and Development，2010，47(10)：1749 – 1755.

[13] 李阳，刘友波，刘俊勇，等. 基于形态距离的日负荷数据自适应稳健聚类算法[J/OL]. 中国电机工程学报：1 – 13 [2019 – 09 – 10]. https://kns.cnki.met/kcms/detaill/1.2107.TM.20171130.1501.001.html.

[14] 程明畅，刘友波，张程嘉，等. 基于分位数半径的动态 K – means 算法[J]. 南京大学学报(自然科学)，2018，54(1)：48 – 55.

作者简介：

林红阳(1972)，高级工程师，主要从事电力市场及能源经济等方面的研究工作。

(收稿日期：2019 – 09 – 19)

(上接第 13 页)

刘俊翔(1986)，硕士研究生，从事电力设备状态评价研究；

王红斌(1972)，硕士研究生，从事电网设备运行研究；

莫文雄(1971)，硕士研究生，从事电网设备运行研究；

彭和平(1981)，硕士研究生，从事电力设备状态评价研究；

王海靖(1987)，硕士研究生，从事电力设备状态评价研究。

(收稿日期：2018 – 08 – 05)